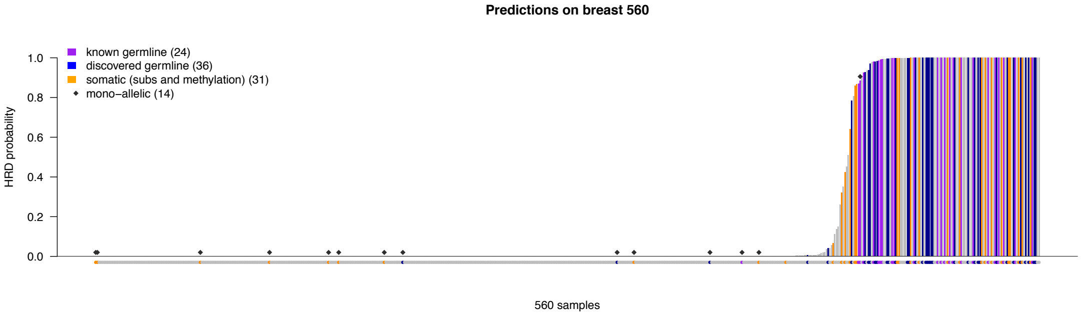
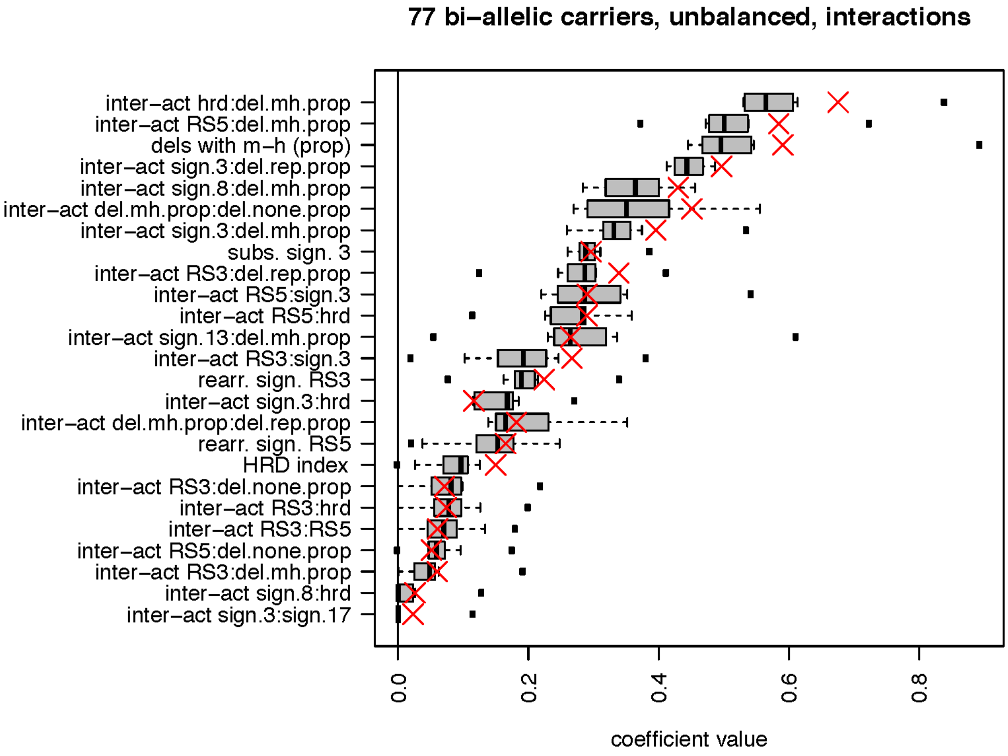
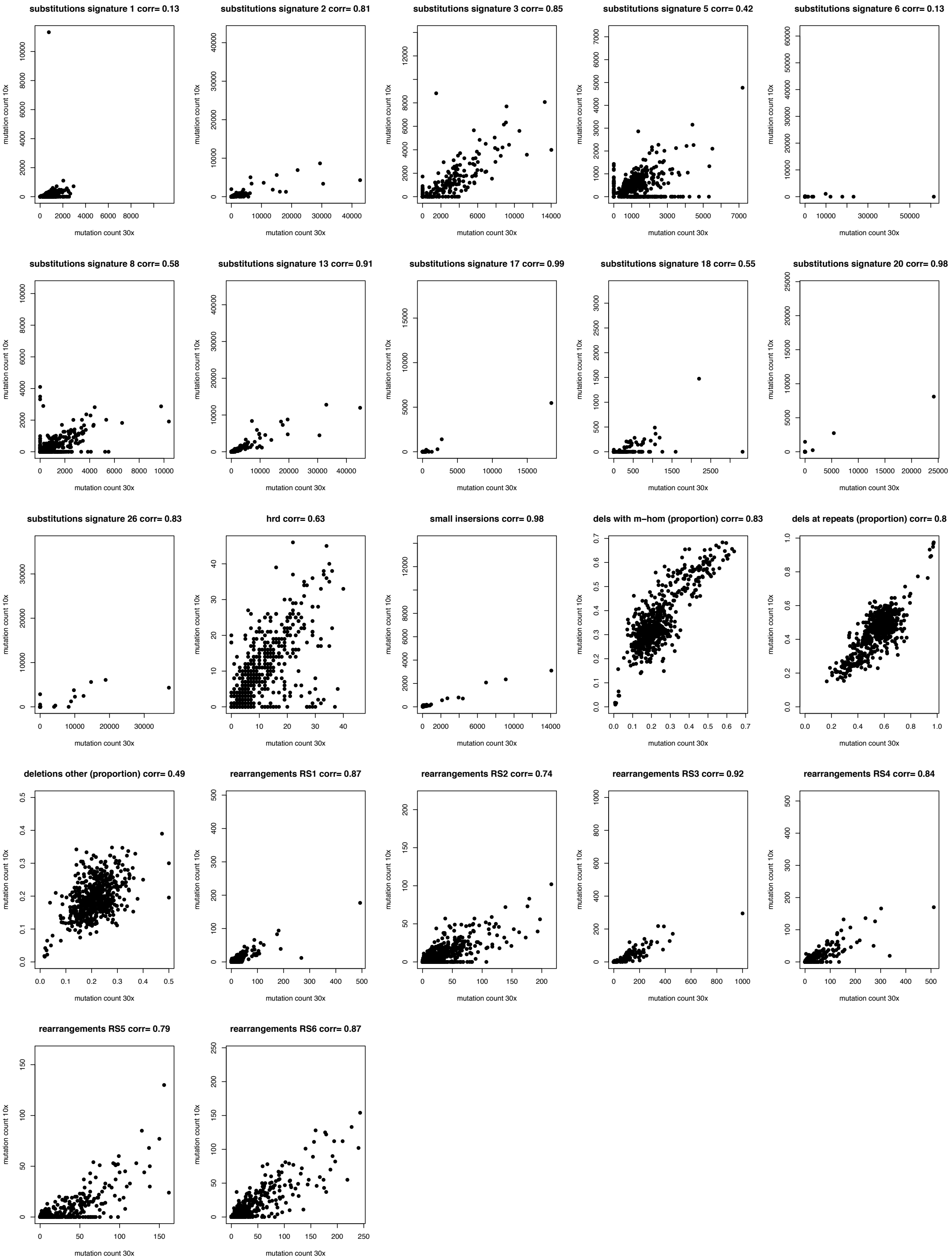
Supplementary Figures



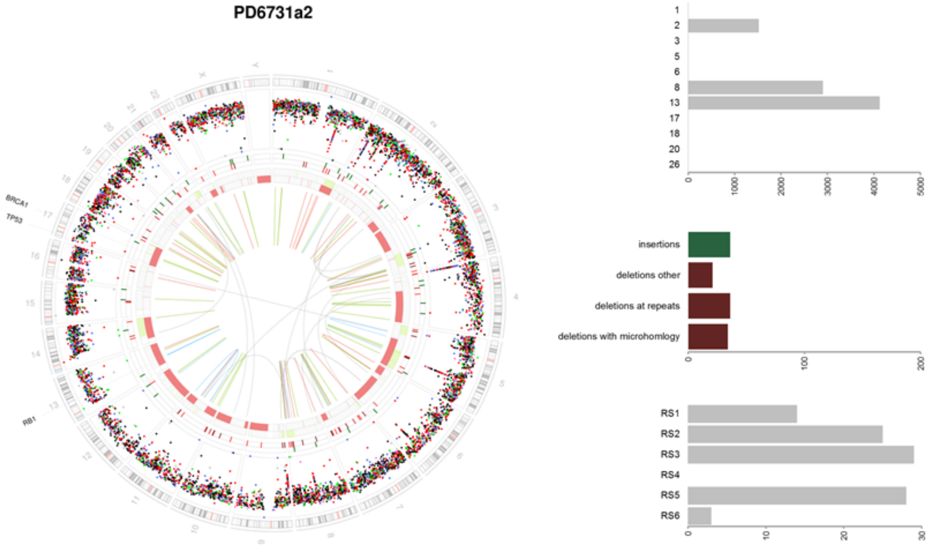
Supplementary Figure 1: Probabilities of *BRCA1/BRCA2* deficiency across entire breast cancer cohort of 560 patients, according to the classifier trained on 22 known germline *BRCA1/BRCA2* mutation carriers (with loss of the other parental allele) and 235 quiescent tumours.



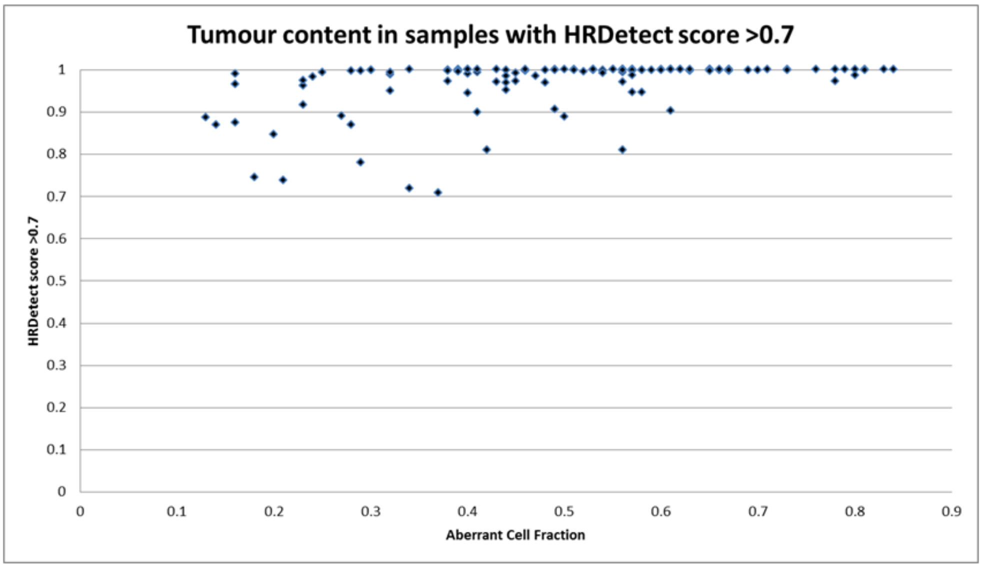
Supplementary Figure 2: Classifier with variable interactions. Box plots of the weights for the genomic features contributing to the classifier. Range of values from 10 replicates of training in cross-validation. HRD=Homologous Recombination Deficiency Index; del = deletions; m-h = microhomology; prop = proportions; subs. = substitutions, sign. = signature, RS = rearrangement signature



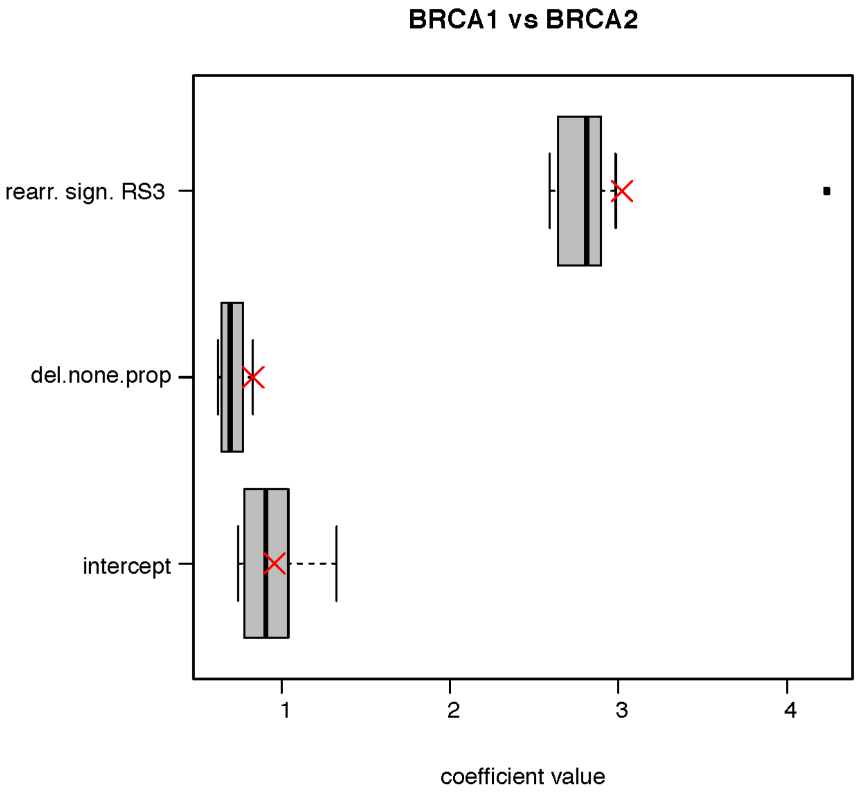
Supplementary Figure 3: Comparison of mutation burden detected in 560 breast cancer samples by high coverage (30x) and simulated low-coverage (10x) whole genome sequencing. Each panel compares mutations counts detected in individual samples, focussing first on single-base substitutions (signatures 1, 2, 3, 5, 6, 7, 13, 17, 18, 20, 26), HRD index, small insertions and deletions, rearrangements (Signatures RS1-RS6).



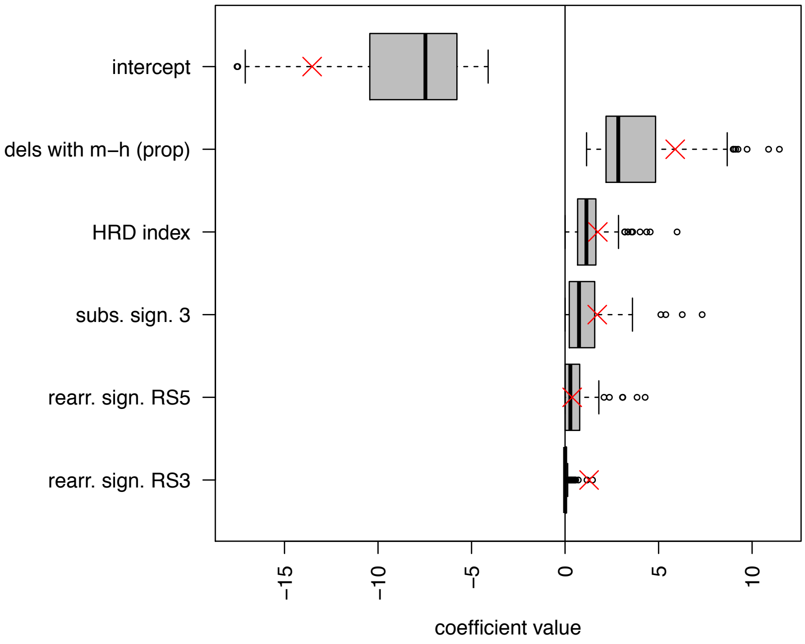
Supplementary Figure 4: Example of a tumour from a patient with a germline *BRCA1* mutation showing biological noise produced by APOBEC related mutations. Left hand side; whole genome plot, right hand side; contribution of signatures from top to bottom: Substitution signatures, indel signatures and rearrangement signatures.



Supplementary Figure 5: HRDetect scores plotted against aberrant cell fraction values produced by ASCAT for the 124 breast cancer samples that scored above the 0.7% threshold.

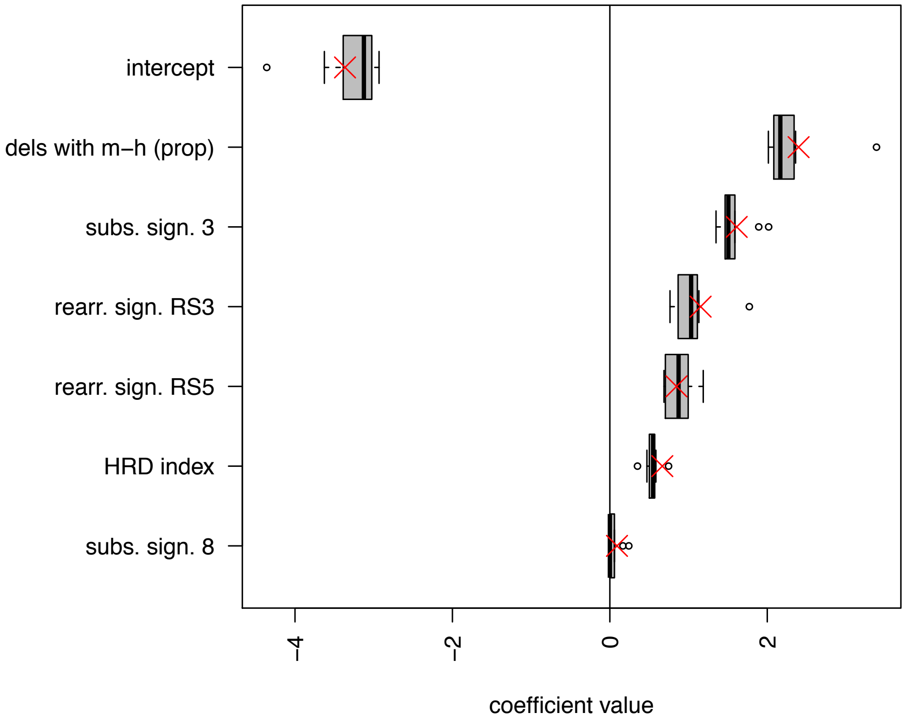


Supplementary Figure 6: Weights for genomic features that allow the predictor to distinguish *BRCA1* tumours from *BRCA2* tumours (using 47 *BRCA1* and 30 *BRCA2* null samples).

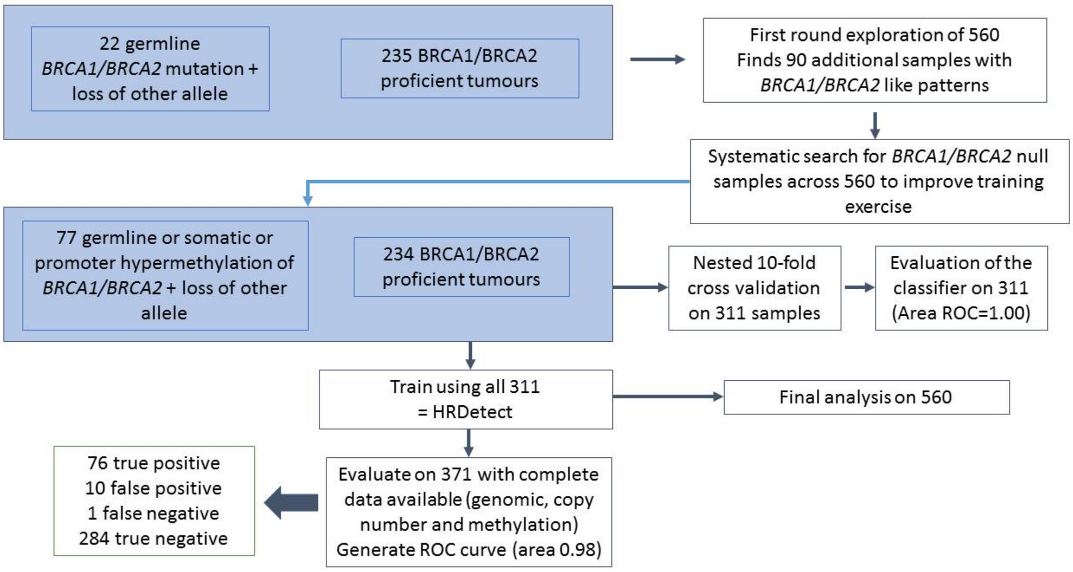


Supplementary Figure 7: Box plots of the weights for the genomic features contributing to the classifier, trained on samples from 22 know germline *BRCA1/2* carriers that are null at the tumour level and 235 control tumours. Range of values from 10 replicates of training in cross-validation are provided in the boxplots. Red crosses indicate the coefficients learnt from all training data.

HRD=Homologous Recombination Deficiency Index; dels = deletions; m-h = microhomology; prop = proportions; subs. = substitutions, sign. = signature, rearr. = rearrangement. In this and subsequent box plots, the midline represents the median, the two edges of the box represent the lower and upper interquartile range (IQR), upper whisker = min(max(x), Q3 + 1.5 × IQR) and lower whisker = max(min(x), Q1 − 1.5 × IQR), and the dots are outliers beyond the whiskers.



Supplementary Figure 8: Boxplots of the weights for the genomic features contributing to the HRDetect classifier. Range of values from 10 replicates of training in cross-validation. Red crosses indicate the final coefficients used in the HRDetect. Genomic features are as described in Supplementary Figure 7.


22 germline *BRCA1/BRCA2* mutation + loss of other allele
235 BRCA1/BRCA2 proficient tumours
First round exploration of 560 Finds 90 additional samples with *BRCA1/BRCA2* like patterns
Systematic search for *BRCA1/BRCA2* null samples across 560 to improve training exercise
77 germline or somatic or promoter hypermethylation of *BRCA1/BRCA2* + loss of other allele
234 BRCA1/BRCA2 proficient tumours
Nested 10-fold cross validation on 311 samples
Evaluation of the classifier on 311 (Area ROC=1.00)
Train using all 311 = HRDetect
Final analysis on 560
76 true positive
10 false positive
1 false negative
284 true negative
Evaluate on 371 with complete data available (genomic, copy number and methylation) Generate ROC curve (area 0.98)

Supplementary Figure 9: Flow diagram showing the steps involved in the training, evaluation and application of the HRDetect predictor. The number of *BRCA1/BRCA2* proficient tumours differs by one sample between the two training rounds because one of the samples with a quiescent genome, PD6042a, was subsequently found to have a biallelic mutation of *BRCA2*. A true positive is a sample that is *BRCA1/BRCA2* null and is correctly given a high HRDetect score (>0.7). A false negative is a sample which has a clear germline or somatic *BRCA1/BRCA2* mutation or promoter hypermethylation and loss of the other allele, but is miscalled with a low HRDetect score. A false positive is a sample that was given a high HRDetect score (>0.7) but in which no *BRCA1/BRCA2* mutation was detected. A true negative is a *BRCA1/BRCA2* proficient tumour correctly given a low HRDetect score.

Supplementary tables 10-14

Genomic feature	Weight
Deletions with micro-homology	5.889
HRD index	1.752
Substitutions signature 3	1.722

Rearrangements RS3	1.285
Rearrangements RS5	0.381

Supplementary table 10: Weights for the genomic features contributing to the classifier. Trained using samples from 22 known germline *BRCA1/BRCA2* carriers and 235 quiescent tumours.

Genomic feature name	Number of times selected as non-zero coefficient (out of 100)
Deletions at micro-homology	100
HRD index	95
Substitutions signature 3	83
Rearrangements RS5	72
Substitutions signature 8	49
Rearrangements RS3	32
Deletions other	7

Supplementary table 11: Stability analysis of genomic features of BRCAness, when trained using half of the data from the total of 22 known germline *BRCA1/BRCA2* carriers and 235 quiescent tumours.

Coefficient name	Number of times non-zero (out of 100)
Deletions at micro-homology	100
HRD index	92
Substitutions signature 3	99
Rearrangements RS5	67
Substitutions signature 8	81
Rearrangements RS3	61
Deletions other	15
Substitutions signature 5	13
Substitutions signature 13	6
Rearrangement signature RS1	2

Supplementary table 12: Stability analysis of genomic features of BRCAness, when trained using half of data from the total of 77 *BRCA1/BRCA2* carriers and 234 quiescent tumours.

Genomic feature	Weight
Deletions with micro-homology	2.398
Substitutions signature 3	1.611

Rearrangements RS3	1.153
Rearrangements RS5	0.847
HRD index	0.667
Substitutions signature 8	0.091

Supplementary table 13: Weights for the genomic features contributing to the classifier. Trained using samples from 77 *BRCA1/ BRCA 2* carriers and 234 quiescent tumours.

Feature	Mean	sd
Rearrangements RS3	1.260	1.657
Rearrangements RS5	1.935	1.483
Substitutions signature 3	2.096	3.555
Substitutions signature 8	4.390	3.179
HRD index (copy-number)	2.195	0.750
Deletions with micro-homology (proportion of all deletions)	0.218	0.090

Supplementary table 14: Mean and standard deviations of genomic features in the cohort of 311 breast cancers used in the final training set

Supplementary Note

1. Variants in *BRCA1* and *BRCA2* and other HR genes

1.1 Variants in BRCA1 and BRCA2

We sought to discover additional germline and somatic mutations of all classes in *BRCA1* and *BRCA2* in the 560 breast cancer cases[1]. Single base substitutions and small insertions/deletions were interrogated using Caveman and Pindel algorithms respectively. While large deletions were investigated using a combination of ASCAT copy number data and rearrangement calls by BRASS. Variants affecting the coding regions of these genes were verified by visual inspection to remove common sequencing artefacts and cross-referenced against dbSNP and ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/) to identify benign polymorphisms. Variants with evidence in both the tumour and corresponding non-neoplastic tissue were deemed to be germline, while those restricted to the tumour were somatic. ASCAT copy number data was used to determine if there had been loss of the alternative allele in the tumour sample. Careful curation of the data was undertaken to avoid misinterpretation of mutation calls. For example, a germline essential splice mutation in PD13418 had accompanying Loss of Heterozygosity (LOH) from ASCAT data. However, visual inspection of the sequencing reads revealed that LOH in the tumour was in favour of the wildtype not the mutant allele. Another example, PD13604 contained two separate somatic *BRCA2* truncating mutations. Although ASCAT copy number analysis indicated there was no LOH of *BRCA2*, there was a potential that the two mutations represented a compound heterozygous inactivation of the gene. However, the overall copy number at the *BRCA2* locus was 4 and read counts suggested that each of the two mutations did not affect all copies of the gene (14-18% of reads). This sample is an APOBEC hypermutator, with over 93,000 substitutions and these somatic mutations may be passenger mutations resulting from this hypermutation.

We used Caveman to collate single base substitutions, which were present in the germline and were not deemed to be deleterious. A total of 127 different germline SNPs and variants of unknown significance (VUS) in *BRCA1* and *BRCA2* were detected in the data set. For the variants present in less than 5% of the samples, ASCAT copy number data was used to determine if there had been loss of the alternative allele in the tumour sample.

Deleterious germline variants, SNPs and VUSs in *BRCA1* and *BRCA2* are included in Supplementary table 4. Deleterious mutations were defined as truncating mutations (nonsense or frame shift), those affecting essential splice sites, whole gene deletions, large intragenic deletions and rearrangements disrupting the gene footprint. In addition, missense mutations which have previously been reported as pathogenic were also considered deleterious. Samples which had biallelic loss of *BRCA1* or *BRCA2* were considered to be *BRCA1/2* null.

1.2 Variants in other HR genes

We searched for deleterious germline variants in the following genes known to be involved in DNA repair via homologous recombination (HR): *ATM, ATR, RAD51C, RAD50, CHEK2, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, FANCM, PALB2 (FANCN), FANCP (BTBD12), ERCC4 (FANCQ)* and *BRIP1*. The following high-penetrance and moderate-penetrance germline susceptibility genes, *TP53, PTEN, ATM, BRIP1, CHEK2, ATR, RAD50, CDH1, PALB2* and *STK11* were also investigated for deleterious germline mutations.

Single base substitutions and insertions/deletions were interrogated using Caveman and Pindel algorithms respectively. We searched for the presence of mutations in both the tumour and non-neoplastic tissue from the same individual, which are normally filtered out of the data when looking for somatic mutations. Any variants affecting the coding regions of these genes were verified by visual inspection to remove common sequencing artefacts and cross-referenced against dbSNP and ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/) to remove benign polymorphisms. Variants with evidence of a deleterious effect on the protein are included in Supplementary table 4. For each variant we used ASCAT copy number data to determine if there had been loss of the alternative allele in the tumour sample.

2. Additional details on Lasso logistic regression modelling and HRDetect

2.1 Exploring the possibility of permitting interactions between genomic covariates

We explored the possibility of permitting interactions between all genomic covariates in order to discover potentially augmented effects of cooperating signatures in our model (Supplementary Figure 2).

Interactions were observed. Other models of logistic regression designed to deal with potentially correlated covariates, including elastic net[10] were explored, which appeared to improve performance during cross-validation. Interactions between genomic features previously associated with BRCAness were detected. The results from the 9:1 cross-validation are demonstrated in Supplementary Figure 2. We then applied these learned weights to generate a ROC curve which did not show any improvement to the performance of the simpler model (HRDetect), where interactions between genomic covariates were not allowed.

2.2 The effect of normal contamination on HRDetect performance

We investigated whether a high proportion of normal DNA contamination (or a reduced tumour cellularity) in a tumour sample had an adverse effect on the ability to detect *BRCA1/BRCA2* deficient samples using HRDetect. Supplementary Figure 5 shows HRDetect scores for the 124 samples that scored above the 0.7 threshold, plotted against aberrant cell fraction values produced by ASCAT (as a measure of tumour content). HRDetect scores consistent with *BRCA1/BRCA2* deficiency were detected across a wide range of tumour content, aberrant cell fraction 0.13 to 0.84, mean 0.499. This was very similar to the aberrant cell fraction for the cohort of 560 samples as a whole, aberrant cell fraction 0.11 to 0.91, mean 0.52.

2.3 HRDetect is able to detect BRCA1/BRCA2 deficiency in the presence of biological noise

Some breast cancers genomes contain a high number of substitution mutations belonging to signatures 2 and 13, which have been attributed to members of the AID/APOBEC family of cytidine deaminases. The presence of such high numbers of mutations can make it more difficult to detect other signatures coexisting in the same sample. Supplementary figure 4 shows an example of a breast cancer from a germline *BRCA1* carrier in which the tumour has a very large number of APOBEC-related signature 2 and 13 mutations. Substitution signature analysis is unable to extract *BRCA1/2* related signature 3 because of the overwhelming presence of signatures 2 and 13. However, HRDetect is still able to recognise *BRCA1* deficiency and produce a high probability score (0.81). It is reassuring to see that because HRDetect uses a combination of signatures from all mutation types, it is able to detect *BRCA1/BRCA2* deficiency despite the presence of biological noise produced by APOBEC-related mutagenesis.

3. Validation of HRDetect in a new cohort of 80 WGS breast cancer

Whole genome sequence (WGS) data from a further cohort of 80 breast cancer were used to validate HRDetect. Internal Review Boards of each participating institution approved collection and use of samples from these 80 patients. Whole genome sequencing as described above was performed on tumours samples to an average depth of 44X and corresponding normal tissue to an average depth of 24X (see Supplementary data table 5 for full details of these samples). Mutation detection, copy number and HRD index calculation and signature analysis were performed as described above. Detection of germline *BRCA1* and *BRCA2* mutations was performed as described above. See Supplementary data table 5 for details of germline mutations along with the contributions of each signature in each sample.

4. Down-sampled 560 breast cancer genomes

To simulate a more economical sequencing strategy involving reduced sequence coverage, sequences from the 560 high coverage (30- to 40-fold) WGS breast cancers and their corresponding normal controls, were randomly sampled to generate low coverage (10-fold, range 9.9 to 10.5, mean 10.0) WGS sequence files for mutation analysis. As expected the number of mutations called in the down sampled data was significantly reduced compared to the full coverage (30- to 40-fold) since many of the mutations failed the usual mutation QC metrics. Indels called by Pindel were particularly affected. Therefore, to improve the yield of mutations a number of the post processing filters, normally employed to ensure mutations are only called in good sequence coverage and with a certain threshold of mutant reads, were relaxed for analysis of the 10-fold coverage genomes (https://github.com/cancerit/cgpPindel/wiki/VcfFilters, filters F016 and F018). As a consequence of relaxing these post-processing filters and also of the reduced coverage of the corresponding normal sequence, the number of germline polymorphisms contaminating the mutation calls, for both the indels and substitutions, was increased. To reduce this polymorphism contamination, the resulting mutation calls were filtered against: dbSNP, 1000 genomes, list of known artefacts, unmatched normal panel, and a set of other genomes and exomes available. Signatures and HRD indices were extracted as described above. Supplementary Figure 3 shows the comparison of each signature produced with full coverage genomes and down-sampled to 10-fold coverage.

5. Application of HRDetect to whole exome sequencing data

To investigate the performance of HRDetect using whole exome sequencing (WES) data we extracted the mutation calls for substitutions and small insertion/deletions situated in the coding region of the genome including splice sites, thus representing the data expected from WES. The data consisted of 43,514 substitutions and 2,787 indels (Supplementary Table 7), representing 1.25% and 0.75% of WGS substitutions and indels respectively. Substitution and indel signatures were extracted from this data as described above and HRDetect weights learnt from whole genome sequencing applied. Not only were the number of mutations significantly reduced but rearrangement signatures and copy number derived HRD index were not available for WES data, which in combination considerably affected the sensitivity of *BRCA1/BRCA2* deficiency detection, falling to 46.8%. We re-trained the algorithm using WES-based data as the input and the sensitivity of the predictor, although still much lower than achieved with WGS data, was improved to 73% at the expense of calling 12 additional samples that were not previously identified as *BRCA1/BRCA2* deficient (Supplementary table 7).

6. Application of HRDetect to other types of cancers

We explored whole genome sequencing data for published samples from the following cancer types; Ovarian cancer, 73 samples[2] and 96 Pancreatic cancer samples from two publications, 60[3] and 36[4] respectively. Raw sequence data were parsed through our somatic-mutation calling pipeline, mutational signatures extracted and copy number profiles obtained as described above. Detection of germline *BRCA1* and *BRCA2* mutations was also performed as described above. See Supplementary table 8 for details of germline mutations along with the contributions of each signature in each sample.

7. Clinical application of HRDetect

To investigate the performance of HRDetect on samples more widely available in the clinic we looked at two different sample sets.

The first sample was DNA extracted from a FFPE (Formalin Fix Paraffin Embedded) tissue sample from a germline *BRCA1* carrier[6]. Whole genome sequencing was performed on DNA extracted from FFPE breast tissue and on normal DNA extracted from blood, mutations called, mutation signatures extracted and HRD index calculated as described above. See Supplementary table 9 for the data from this sample.

The second sample set contained 18 samples from nine patients who had been treated with neoadjuvant anthracyclines +/- taxanes[5]. DNA extracted from pre-treatment needle biopsy samples was available for all nine patients; in addition, five samples had two separate pre-treatment biopsies, including one patient with multifocal tumours. For four of the five cases that showed residual disease following treatment DNA was also available from larger tissue blocks obtained at post-treatment surgery. See Supplementary table 9 for details of the samples. One patient, (PD9773), with extremely low tumour cellularity in both biopsies and with hardly any mutations, was excluded from this analysis. Again, whole genome sequencing was performed on DNA extracted from the tumour sample together with corresponding normal tissue, mutations called, mutation signatures extracted and HRD index calculated as described above and used as input for HRDetect.

8. Distinguishing *BRCA1* from *BRCA2* tumours

Although tumours deficient in *BRCA1* and *BRCA2* share many genomic features, we observed that *BRCA1* deficient tumours tended to have an excess of Rearrangement Signature 3 (characterised by small tandem duplications), while *BRCA2* deficient tumours have a higher proportion of rearrangement Signature 5 (non-clustered deletions <100 kb). In readiness for any future therapeutic intervention which might require distinguishing *BRCA1* and *BRCA2* tumours from each other, we looked at training the predictor to the distinguish tumours with defects in the two genes. We trained the predictor in a similar way to described above using 47 samples with *BRCA1* mutations plus loss of the alternative allele as the positive sample set and 30 samples with *BRCA2* mutations plus loss of the alternative allele as the negative sample set. Two genomic features, rearrangement Signature 3 and the proportion of deletions without distinctive junctional characteristics were found to separate *BRCA1* from *BRCA2* tumours (Supplementary Figure 6). At present to apply this to a general breast cancer sample set would require two separate predictor steps; the first to identify tumours that were *BRCA1*/*BRCA2* deficient and the second to distinguish *BRCA1* from *BRCA2* deficient tumours.

References:

1. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54 (2016).
2. Patch, A.M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489-94 (2015).
3. Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495-501 (2015).
4. Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47-52 (2016).
5. Yates, L.R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med* **21**, 751-9 (2015).
6. Yates, L.R. *et al*. Genomic evolution of breast cancer metastasis and relapse. *Manuscript submitted.*